DOCUMENT RESUME

ED 385 577 TM 024 015

AUTHOR Wainer, Howard; Thissen, David
TITLE Choosing: A Test. ETS Program Statistics Research.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-92-67; ETS-TR-92-25
PUB DATE Nov 92
NOTE 15p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Adaptive Testing; Cost Effectiveness; Responses; *Scoring; *Selection; *Test Items; Test Use
IDENTIFIERS *Choice Behavior; National Assessment of Educational Progress

ABSTRACT

If examinees are permitted to choose to answer a subset of the questions on a test, just knowing which questions were chosen can provide a measure of proficiency that may be as reliable as would have been obtained from the test graded traditionally. This new method of scoring is much less time consuming and expensive for both the examinee and the testing organization. Moreover, because of the decreased response burden, it may be expected that its use may reduce the nonresponse rate in such low impact educational assessments as the National Assessment of Educational Progress (NAEP). It is recommended that in assessments that allow the examinee to choose among a set of items, an attempt be made to encourage the examinee to record the item that will be answered before they are actually attempted. In this way, it may be possible to obtain much of the information that would have been contained in the actual responses, even if the examinee chooses finally not to answer. Four tables illustrate the discussion, and a technical appendix discusses calculating the reliability coefficient. (Contains six references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RR-92-67

ED 385 577

Choosing: A Test

Howard Wainer
Educational Testing Service

David Thissen
University of North Carolina at Chapel Hill

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

H. I. BRAUN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

PROGRAM STATISTICS RESEARCH

TECHNICAL REPORT NO. 92-25

Educational Testing Service
Princeton, New Jersey 08541

TM024015

BEST COPY AVAILABLE

The Program Statistics Research Technical Report Series is designed to make the working papers of the Research Statistics Group at Educational Testing Service generally available. The series consists of reports by the members of the Research Statistics Group as well as their external and visiting statistical consultants.

Reproduction of any portion of a Program Statistics Research Technical Report requires the written consent of the author(s).

Choosing: A Test

Howard Wainer
Educational Testing Service

David Thissen
University of North Carolina at Chapel Hill

Program Statistics Research
Technical Report No. 92-25

Research Report No. 92-67

Educational Testing Service
Princeton, New Jersey 08541

November 1992

Copyright © 1992 by Educational Testing Service. All rights reserved.

Choosing: A test[1]

Howard Wainer
Educational Testing Service

David Thissen
University of North Carolina
at Chapel Hill

Abstract

If examinees are permitted to choose to answer a subset of the questions on a test, by knowing only which questions were chosen we can obtain a measure of proficiency that may be as reliable as would have been obtained from the test graded traditionally. This new method of scoring is much less time consuming and expensive for both the examinee and the testing organization. Moreover, because of the decreased response burden we would expect that its use may reduce the nonresponse rate in such low impact educational assessments as the NAEP. We recommend that in assessments that allow the examinee to choose among a set of items, an attempt be made to encourage the examinee to record the item that will be answered before they are actually attempted. In this way we may be able to obtain much of the information that would have been contained in the actual responses, even if the examinee chooses finally not to answer.

[1]This work was supported through funds provided in a contract by the research program of the Graduate Record Examination to the first author and we are pleased to be able to acknowledge their generosity. We would also like to thank Nancy Allen, Charles Lewis, Gene Johnson, Robert Mislevy, Linda Steinberg and Xiang-bo Wang for their help and advice.

Introduction

Much of human life is defined by choices made: which college to attend; which job to accept; if, who, and when to marry; if, when, and how many children to have. Hence what choices a person makes often provide an enormous amount of information about that person. Indeed what one chooses to do can sometimes be more informative about a person's ability than how well that task is subsequently done. Most doctoral candidates, with a bit of introspection, would agree that the most difficult aspect of a dissertation is choosing the topic. An unwise choice may lead nowhere, may not be substantial enough to qualify, or may be too difficult to be completed in a timely way.

Until recently modern testing has largely eschewed allowing examinees to choose which items in a test they would like to answer. Gulliksen (1950, p. 337-338), in his now-classic treatise, warned strenuously against the practice because of the unrealistic assumptions required to equate the different forms thus created. Equating is required by the canons of testing to assure the fairness of the test (American Psychological Association, 1985). Over the last decade or two, the use of choice has started to creep back into large-scale testing (see Pomplun et al, 1992 for a review of this practice within the Advanced Placement exam programs of the College Board).

Examinee item-choice is now allowed in a substantial number of tests, including the writing portion of the National Assessment of Educational Progress (Johnson & Zwick, 1988). Equating the test forms constructed through examinee choice is not a great deal easier now than it was 40 years ago when Gulliksen warned against it; however, using a strong statistical model, one can be explicit about the assumptions required to accomplish equating and, after making these assumptions, obtain both equated parameters for all items and proficiency estimates for all examinees (Wainer, Wang & Thissen, 1991).

The estimates obtained from a model such as that used by Wainer et al (1991) are based on actually scoring the items. How much information can be obtained by noting only which items were chosen? If choice requires wisdom and knowledge, then we should be able to measure how much wisdom and knowledge are being displayed by examinees merely from observing which items they choose to answer. Or, to move to a more subjunctive mood, by observing which items they would choose to answer.

In the next section of this paper we describe the results of a small study that considers this issue. We conclude that on the College Board's Advanced Placement Chemistry Test we obtain less information from the choice of items than we do from examinees choosing the items, answering them, and having them scored by expert judges. Since item choice is a subset of the information available for the latter scheme this is not surprising. However, we show how a small modification in the number of choices made would allow substantial increases in information.

Example: AP Chemistry

The data we use in the illustration are from the 1989 Advanced Placement Examination in Chemistry. A full description of this test, the examinee population, and the scoring model is found in Wainer, Wang & Thissen (1991) and will not be fully repeated here. For the purposes of this study we consider only the five constructed response items in Part II, Section D. Section D has five problems (Problems 5, 6, 7, 8 and 9) of which the examinee must answer three. This section accounts for 19% of the total grade.

This form of the exam was taken by approximately 18 thousand students[2] in 1989. The test form has been released and interested readers may obtain copies of it with the answers and a full description of the scoring methodology from the College Board. These items are problems that are scored on a scale from zero to eight. The scoring scheme is analytic and quite rigorous.

Since examinees had to answer three out of the five questions, a total of 10 choice groups were formed with each group taking a somewhat different test form than the others. Each group had at least one problem in common with every other group and we used this overlap to place all examinee selected forms onto a common scale. We did this by fitting a polytomous IRT model to all ten forms simultaneously (see Wainer, Wang & Thissen, 1991 for details). As part of this fitting procedure we obtained estimates of the mean value of each choice group's proficiency (μ_i) as well as the marginal reliability of this section of the test. Our findings are summarized in Table 1.

[2]The actual number of examinees was 18,466; however 31 tests were handed-in essentially blank and so are excluded from the analysis.

Table 1

Group	*Problems Chosen*	*Mean Group Proficiency (μ_i)*	*n*	*Cronbach's α*
1	567	-1.02	2,555	0.63
2	679	-0.04	121	0.65
3	568	0.00*	5,227	0.57
4	579	0.04	753	0.64
5	578	0.08	4,918	0.51
6	678	0.08	1,392	0.54
7	569	0.09	457	0.67
8	689	0.40	407	0.57
9	789	0.43	898	0.59
10	589	0.47	1,707	0.59

*The mean for Group 3, the largest group, is fixed at 0.0 to set the location of the proficiency scale.

The proficiency scale used had a standard deviation of one; thus those examinees who chose the first three items (5, 6, and 7) showed considerably less proficiency than any other group. Next the groups labeled 2 through 7 were essentially indistinguishable in performance from one another. Last, Groups 8, 9 and 10 were the best performing groups. The reliability of the ten overlapping forms of this three item test ranged from .51 to .67 with a mean of .57.

Suppose we think of Section D as a single 'item' with an examinee falling into one of ten possible categories, and the estimated proficiency of each examinee is the mean score of everyone in their category. How reliable is this one item test? We can calculate an analog of reliability, the squared correlation of proficiency (θ) with estimated proficiency ($\hat{\theta}$) from the between group variance [var(μ_i)] and the within group variance (which is here fixed at unity). This index of reliability

$$\{r^2(\hat{\theta}, \theta) = \text{var}(\mu_i)/[\text{var}(\mu_i) + 1]\}$$

is easily calculated. See the technical appendix for a derivation. The variance of the μ_i is .17, and so $r^2(\hat{\theta}, \theta)$ is .15 (= .17/1.17). While .15 is less than .57, it is also larger than zero, and it is easier to obtain.

Consider a 'choice test:' it comprises 'items' like Section D, except that instead of asking examinees to pick three of the five questions and answer them, we ask them to indicate which three questions they would answer, and then go on to the next 'item'. This takes less time for the examinee, and is far easier to score.

The choice task of picking three out of five questions to answer is a single testlet. We calculate the reliability of a test made up of any number of such testlets in the same way that we do with any other test, through the Spearman-Brown Prophesy formula. Thus, if we ask the examinee to pick three from five on one set of topics and then three from five on another we have effectively doubled the test's length and its reliability has rises from .15 to .26. The estimated reliabilities for tests built of various numbers of such choice testlets are shown in Table 2.

While time constraints under most circumstances prevent asking examinees to answer twenty or thirty long items, it may be plausible to ask examinees to consider the questions and indicate which ones they would answer if they were asked to do so. A computer-administered test might even select a couple for them to actually answer, to 'keep them honest.'

Table 2

Spearman-Brown Extrapolation
For Building a Test of Specified Reliability

Number of Testlets*	Reliability
1	0.15
2	0.26
3	0.35
4	0.41
5	0.47
10	0.64
20	0.78
30	0.84
50	0.90

* Here, each testlet comprises the task of selecting three questions out of five.

Measured in bits of information, a choice test in which an examinee decides which 3 of 5 items to answer may be thought of as equivalent to 3.3 binary items ($2^{3.3}$ = 10 score groups). The binary 'question' is "Would you answer this question?" In a substantial stretch of imagination, we might apply the Spearman-Brown formula somewhat differently, to predict the reliability of different kinds of tests. Suppose, for example, we ask the examinees to report for each of eight questions whether they would answer. This procedure would divide the examinees into 2^8=256 groups, and if the

Spearman Brown formula applies we would expect the reliability to increase to about .3. Table 3 shows Spearman-Brown extrapolated reliabilities for choice tests equivelent to varying numbers of binary items.

Table 3

Spearman-Brown Extrapolation
for building a choice test of specified reliability
"Test Length" is measured in binary choices

Test Length	Reliability
2	0.1
5	0.2
8	0.3
12	0.4
19	0.5
28	0.6
44	0.7
75	0.8
168	0.9

The first column of Table 3 shows the number of binary choices the Spearman-Brown formula requires to yield the reliability in the second column. This suggests that it may be possible to obtain a test whose reliability is of the order of .6 if we present 28 questions, and asked the examinees to indicate which they would answer if they were required to do so. This would yield a test of about the same reliability as the current one.

The values in Table 3 are based on items to be chosen-among that have the properties of those comprising Section D for the 1989 AP Chemistry examination: The items must vary substantially in difficulty, and this must be more or less apparent to the examinees, depending on their proficiency. In addition, some mechanism would be required (in practice) to prevent the examinees from responding (or being coached to respond) by simply 'choosing' all of the items! As before, in a computer-administered version of a 'choice' test, this potential problem could be solved (and additional information acquired) by using an algorithm that required the examinee to actually answer one or more of the chosen problems.

Conclusion & Discussion

How much information is obtained by requiring examinees to actually answer questions, and then grading them? This effort has some reward for the AP Chemistry test

where the constructed response section is analytically scored. But how much are these rewards diminished when the test's scoring schema are less rigorous?

Shown in Table 4 are the reliabilities for the constructed response sections of all Advanced Placement Tests. Note that there is very little overlap between the distributions of reliability for analytically and holistically scored tests, the latter being considerably less reliable. Chemistry is a little better than average among analytically scored tests with a reliability of .78 for its constructed response sections.

Table 4

Reliabilities of Constructed Response sections of AP Tests

Analytically scored	Score Reliability	Holistically Scored
Calculus AB	0.85	
Physics B	0.84	
Computer Science	0.82	
Calculus BC	0.80	
French Language	0.79	
Chemistry	0.78	
Latin-Virgil	0.77	
Latin—Catullus-Horace	0.76	
Physics C — Electricity	0.74	
Music Theory, Biology	0.73	
Spanish Language	0.72	
Physics C — Mechanics	0.70	History of Art
	0.69	French Literature
	0.63	Spanish Literature
	0.60	English Language & Composition
	0.56	English Literature & Composition
German Language	0.50	
	0.49	American History
	0.48	European History
	0.29	Music: Listening & Literature

It is sobering to consider how a test that uses only the options chosen would compare to a less reliable test (i.e., any of the holistically scored tests). The structure of such a choice test might be to offer a set of say, ten candidate essay topics, and ask the examinee to choose those topics that he/she would like to write on. And then stop.

It is not our intention to suggest that it is better to have examinees choose questions to answer than it is to actually have them answer them.[3] We observe only that if the purpose is measurement, a great deal of information can be obtained from the choices made. Moreover, one should feel cautioned if the test administration and scoring schema that one is using yields a measuring instrument of about the same accuracy that would have been obtained ignoring the performance of the examinee entirely.

We have used estimates of proficiency for each choice group that were obtained using expert scoring for individuals who actually answered the questions. This guarantees that the choice test is measuring the same latent dimension as the traditional test. While such information might, in practice, be obtained in pretesting, it does represent a significant effort. Pretesting also may cause some security problems.

However, if one is willing to sacrifice the immediate connection between the choice scores and the scores on those specific problems, there is another way. Different test forms generated by examinee choices are often equated through a section of the test in which it is mandatory to answer all questions. Such an equating is valid if the test is unidimensional. Quite often this assumption is valid (Thissen, Wainer & Wang, 1992) and the mandatory section is built of items that can be objectively scored. Using such a section to obtain estimates of the choice group means would serve to calibrate the choice test. In formal terms we would then fit an appropriate IRT model to the objectively scored mandatory items and also to the choice items scored, say, 1 if chosen and 0 if not. The estimates of proficiency obtained would be analogous to those we have portrayed here. The marginal gain in the test's precision due to the choice items would also be analogous to the results presented above. We predict that such a test can contain the same items currently used, be completed in less time, and be scored far more quickly and economically. Moreover, for some tests, it may yield greater precision.

It may be worth considering a 'choice test' in large scale assessments (like NAEP) in which there is considerable nonresponse to constructed response items, ostensibly because of their extra response load. A 'choice test' can ask the same questions, but not require as much work from the examinee and so perhaps generate a higher response rate. At the very least we can elicit the choices and then ask the examinees to actually answer the items. We then have some information on those examinees who opt not to answer the questions, but who do indicate what their choices would have been.

---

[3] Although our cynical colleague Nick Longford commented that this "Suits perfectly the current American culture in which no one ever actually does anything, but is concerned instead with management."

References

American Psychological Association. (1985). *Standards for educational and psychological testing*. Washington, DC: Author.

Gulliksen, H. O. (1950). *Theory of mental tests*, New York: Wiley. (Reprinted in 1987 by Lawrence Erlbaum Associates; Hillsdale, NJ).

Johnson, E. G., & Zwick, R. J. (1988). *The NAEP Technical Report.* Princeton, NJ.: Educational Testing Service.

Pomplun, M., Morgan, R., & Nellikunnel, A. (1992). *Choice in Advanced Placement Tests.* ETS Statistical Report (SR -92-51): Educational Testing Service, Princeton, NJ.

Thissen, D., Wainer, H. & Wang, X-B. (1992). *How unidimensional are tests comprising both Multiple-choice and Free-Response Items? An analysis of two tests.* ETS Technical Report (92-xx). Princeton, NJ.: Educational Testing Service.

Wainer, H., Wang, X. B., & Thissen, D. (1991). *How well can we equate test forms that are constructed by examinees?* ETS Technical Report (RR -91-57): Educational Testing Service, Princeton, NJ.

Technical Appendix[4]

How can we calculate a reliability coefficient from the classification of examinees by their choice of items? Let us assume that we know the mean proficiency of all examinees in each choice group. We will index examinees by j and choice groups by i, and the model we use is

$$\theta_{ij} = \mu_i + z_{ij} \qquad \text{(A1)}$$

where the proficiency of person j in group i is θ_{ij} and is distributed normally with mean μ_i and variance 1. We represent the deviation of each person j within group i from that group's mean as z_{ij}.

If we estimate $\hat{\theta}_{ij}$ with μ_i, the mean of group i, that is

$$\hat{\theta}_{ij} = \mu_i \qquad \text{(A2)}$$

the correlation between θ_{ij} and $\hat{\theta}_{ij}$ is analogous to validity if we think of θ_{ij} as the analog of true score. The square of this correlation can be thought of as a measure of reliability. Keeping this in mind, we can derive a computational formula for

$$r^2(\hat{\theta}_{ij}, \theta_{ij})$$

by noting that

$$r^2(\hat{\theta}_{ij}, \theta_{ij}) = [\,\mathrm{cov}(\hat{\theta}_{ij}, \theta_{ij})]^2 / [\mathrm{Var}(\theta_{ij}) \times \mathrm{Var}(\hat{\theta}_{ij})]. \qquad \text{(A3)}$$

In the numerator,

$$\begin{aligned}\mathrm{cov}(\hat{\theta}_{ij}, \theta_{ij}) &= \mathrm{cov}\,[E(\theta_{ij} \mid i), E(\hat{\theta}_{ij} \mid i)] + E[\mathrm{Cov}(\hat{\theta}_{ij}, \theta_{ij} \mid i)] \\ &= \mathrm{cov}(\mu_i, \mu_i) + E[\mathrm{Cov}(\theta_{ij}, \mu_i \mid i)] \\ &= \mathrm{var}(\mu_i)\end{aligned}$$

The rightmost term in the initial expression $\{E[\mathrm{Cov}(\hat{\theta}_{ij}, \theta_{ij} \mid i)]\}$ is zero and hence the expression reduces to the covariance of μ_i with itself, or the variance of μ_i.

This is the expression in numerator that we need to compute (A3). The denominator requires the variance of both $\hat{\theta}_{ij}$ and θ_{ij}. These are easily computed from

$$\begin{aligned}\mathrm{Var}(\theta_{ij}) &= \mathrm{Var}\,[E(\theta_{ij} \mid i)] + E[\mathrm{Var}(\theta_{ij} \mid i)] \\ &= \mathrm{var}(\mu_i) + 1, \qquad \text{(A4)}\end{aligned}$$

and

---

[4]We are grateful to Charles Lewis who suggested this analog for reliability, provided a derivation, and cautioned against its too broad usage.

$$\text{Var}\,(\hat{\theta}_{ij}) = \text{var}(\mu_i)\,. \qquad \text{(A5)}$$

Substituting these results into (A3) yields

$$r^2(\hat{\theta}_{ij},\ \theta_{ij}) = \text{var}(\mu_i)/[\text{var}(\mu_i) + 1]. \qquad \text{(A6)}$$

The estimate of $\text{var}(\mu_i)$ we obtained from Section D of AP Chemistry is .17 and hence the estimated reliability [from (A6)] is .15.